CS 124

MATRIX DECOMPOSITIONS AND STATISTICAL CALCULATIONS

BY

GENE H. GOLUB

TECHNICAL REPORT NO. CS 124
MARCH 10, 1969

COMPUTER SCIENCE DEPARTMENT
School of Humanities and Sciences
STANFORD UNIVERSITY

Matrix Decompositions

and

Statistical Calculations*

BY

Gene H. Golub

Computer Science Department

Stanford University

*Invited paper to be presented at the Conference on Statistical Computation, University of Wisconsin, Madison, Wisconsin, April 28-30, 1969. This work was in part supported by the National Science Foundation and Office of Naval Research.

ABSTRACT

Several matrix decompositions which are of some interest in statistical calculations are presented. An accurate method for calculating the canonical correlation is given.

Table of Contents

0. Introduction

With the advent of modern digital computers, many of the well known hand calculator methods for making statistical calculations have been revised. For example, Hotelling [19] proposed a number of methods for solving matrix problems. Yet today almost none of these methods are in current use. In this paper, we shall present several well known matrix decompositions and show their relevance to statistical calculation. Some of the properties of the numerical algorithms shall be discussed.

1. Cholesky decomposition

Let A be a real, symmetric, positive definite matrix of order n. It is well known that we may factor A so that

$$A = R^T R \tag{1.1}$$

where R is an upper triangular matrix (▽). The decomposition (1.1) is known as the Cholesky decomposition. The calculation of R may be performed in two ways.

a) Complete Cholesky Decomposition Algorithm (CCDA)

Let

$$r_{11} = (a_{11})^{1/2} \quad \text{and} \quad r_{1j} = a_{1j}/r_{11} \; (j=2,\ldots,n) .$$

Then for $i = 2,\ldots, n$,

$$\left.\begin{aligned} r_{ii} &= \Big(a_{ii} - \sum_{k=1}^{i-1} r_{ki}^2\Big)^{1/2} , \\ r_{ij} &= \Big(a_{ij} - \sum_{k=1}^{i-1} r_{ki} r_{kj}\Big)/r_{ii} \; (j=i+1,\ldots, \bullet \quad \blacksquare . \end{aligned}\right\} \tag{1.1}$$

b) Sequential Cholesky Decomposition Algorithm (SCDA)

Let

$$a_{ij}^{(1)} = a_{ij} .$$

Then for $k = 1,2,\ldots, n$,

$$\left.\begin{aligned} r_{kk} &= (a_{kk}^{(k)})\ ,\ r_{kj} = a_{kj}^{(k)}/r_{kk}\ ,\ (j>k)\ , \\ a_{ij}^{(k+1)} &= a_{ij}^{(k)} - \frac{a_{ki}^{(k)} a_{kj}^{(k)}}{a_{kk}^{(k)}}\ ,\ i\ ,\ j = k+1, \ldots,\ n. \end{aligned}\right\} \qquad (1.2)$$

Since the Cholesky decomposition is unique when $r_{ii}>0$, each of the algorithms produces the same R . Each of these methods require $n^3/6 + O(n^2)$ multiplications plus n square roots. The CCDA has the advantage that if the i^{th} row of the matrix R is being computed, then it is only necessary to have available the i^{th} row of the matrix A , and the (i-1) previously computed rows of R . This is especially advantageous when the matrix A is so large that it is necessary to store it in auxilliary storage.

2. Accuracy of the Cholesky decomposition

J. Wilkinson [31] has given an error analysis of the SCCA. He assumes that the error in the basic operations are as follows:

$$fl(a \pm b) = a(1+\varepsilon_1) \pm b(1+\varepsilon_2)$$
$$fl(a \times b) = ab(1+\varepsilon_3) \qquad |\varepsilon_i| \leq 2^{-t}$$
$$fl(a/b) = (a/b)(1+\varepsilon_4)$$

where a mantissa of t binary digits is used. The notation $fl(a \; op \; b)$ indicates the result of the operation with two floating point numbers a and b when standard floating point arithmetic is used. Furthermore, it is assumed that if

$$x = fl(\text{sqrt } a)$$

then

$$x^2 = a(1+\varepsilon) \text{ with } |\varepsilon| \leq 2 \times 2^{-t_1}$$

where

$$\boxed{t_1 = t - \log_2(1.06)} \quad .$$

When the SCDA is used, arithmetic errors are introduced at each stage of the calculation. Indeed it is possible that for some positive definite matrices it will be impossible to complete the algorithm because

of the roundoff error. Wilkinson has given sufficient conditions for which it is possible to complete the SCDA. In addition, he has shown that the computed Cholesky decomposition will be exact for some perturbed matrix $A + E$. Let $\|...\|_2$ indicate the spectral norm, and let $\bar{R}$ be the computed Cholesky factor.

Theorem 1. (Wilkinson): If A is a positive definite matrix of order $n > 10$, then provided

$$\lambda_{min}(A) \geq 20n^{3/2}2^{-t}\|A\|_2$$

the Cholesky factor $\bar{R}$ can be computed without breakdown and the computed $\bar{R}$ satisfies the relation

$$\bar{R}^T\bar{R} = A + E ,$$

$$\|E\|_2 \leq 2.5n^{3/2}2^{-t}\|A\|_2 .$$

Thus the relative perturbation viz $\|E\|_2/\|A\|_2$, is but a few units of the mantissa for the Cholesky factorization. The above result is independent of the choice of pivots.

3. Solution of linear equations

Given the Cholesky decomposition, it is a simple matter to solve a system of linear equations or to compute the inverse. To solve $A\underset{\sim}{x} = \underset{\sim}{b}$, the most convenient procedure is to first solve

$$R^T\underset{\sim}{y} = \underset{\sim}{b} \tag{3.1}$$

and then solve

$$R\underset{\sim}{x} = \underset{\sim}{y} \ .$$

Since R is $\triangleleft$ and R^T is $\triangleright$, this requires a total of $n^2 + O(n)$ multiplications. To compute the inverse, compute R^{-1} which is $\triangleleft$ and then compute $R^{-1}R^{-T}$, taking advantage of the triangular form of R^{-1} and the symmetry of A^{-1}; this requires $\frac{n^2}{3} + O(n^2)$ multiplications. Thus to invert a positive definite matrix requires $n^2/2 + O(n^2)$ multiplications which is fewer multiplications than multiplying two matrices by the usual algorithm!

Because of the roundoff error, equations (3.1) and (3.2) can be replaced by a perturbed system of equations. Thus, in reality we have

$$(\bar{R}^T+\delta\bar{R}^T)\underset{\sim}{u} = \underset{\sim}{b}$$

and

$$(\bar{R}+\delta\bar{R})\underset{\sim}{z} = \underset{\sim}{u} \ ,$$

and this implies that

$$(A+\delta A)\underset{\sim}{z} = \underset{\sim}{b}$$

where

$$\delta A = E + \delta\bar{R}^{T}\times \bar{R} + \bar{R}^{-T} \times \delta\bar{R} + \delta\bar{R}^{T} \times \delta\bar{R} \ .$$

Using Theorem 1 and Wilkinson's bounds for solving triangular systems [30 , pg 99], it can be shown that

$$\frac{\|\delta A\|_2}{\|A\|_2} \leq 5n^{3/2}2^{-t_1} \qquad (3.3)$$

when the conditions of Theorem 1 are satisfied.

The bound given by (3.3) is quite gross but it does indicate that solving equations using the Cholesky decomposition leads to a relatively small perturbation in the original data. Note also that we can determine a bound on the residual vector $\underset{\sim}{r} = \underset{\sim}{b} - A\underset{\sim}{z}$. Since $(A+\delta A)\underset{\sim}{z} = \underset{\sim}{b}$, $\underset{\sim}{r} = \delta A\underset{\sim}{z}$ and thus

$$\|\underset{\sim}{r}\|_2 \leq 5n^{3/2}2^{-t_1}\|A\|_2 \ .$$

Of course, if the norm of the residual vector is small it is not true an accurate solution has been determined since

$$x - z = A^{-1}\underset{\sim}{r}$$

and hence

$$\|\underset{\sim}{x}-\underset{\sim}{z}\|_2 \leq \|A^{-1}\|_2\|\underset{\sim}{r}\|_2 .$$

It is possible to bound the norm of the relative error providing an upper bound for the condition number $\kappa(A) = \|A\|_2\|A^{-1}\|_2$ is known. Since $\underset{\sim}{z} - \underset{\sim}{x} = -A^{-1}\delta A\underset{\sim}{z}$, a short manipulation shows that when $\|\delta A\|_2\|A^{-1}\|_2 < 1$,

$$\frac{\|\underset{\sim}{z}-\underset{\sim}{x}\|_2}{\|\underset{\sim}{x}\|_2} \leq \frac{\rho\kappa(A)}{1-\rho\kappa(A)} \qquad (3.4)$$

where $\rho = \|\delta A\|_2/\|A\|_2$. This bound is independent of the method used.

4. Conditioning of matrices

Since the bound (3.4) is dependent upon the 'condition number, it is frequently desirable to replace the original system of equations $A\underset{\sim}{x} = \underset{\sim}{b}$ by a new system

$$DAD\underset{\sim}{w} = D\underset{\sim}{b}$$

where D is a diagonal matrix with non-zero diagonal elements. Let $\mathcal{D}_n$ be the set of all n x n diagonal matrices with non-zero diagonal elements. We wish to choose $\hat{D}$ so that

$$\kappa(\hat{D}A\hat{D}) < \kappa(DAD) \text{ for all } D\varepsilon\mathcal{D}_n \ .$$

A symmetric matrix is said to have Property A if there exists a permutation matrix II such that

$$\Pi^T A \Pi = \left(\begin{array}{c|c} \Delta_1 & S \\ \hline S^T & \Delta_2 \end{array}\right)$$

where $A_1\varepsilon\mathcal{D}_p$ and $\Delta_2\varepsilon\mathcal{D}_q$ and $p+q=n$. All tri-diagonal matrices have Property A.

Let $\mathring{D}\varepsilon\mathcal{D}_n$ and $\{\mathring{D}\}_{ii} = 1/\sqrt{a_{ii}}$. Forsythe and Straus [10] have shown that for matrices that possess Property A, $\hat{D} = \mathring{D}$. More generally, for all positive definite matrices A, van der Sluis [29] has shown that

$$\kappa(\mathring{D}A\mathring{D}) < n\kappa(\hat{D}A\hat{D}) \ . \qquad (4.1)$$

Therefore in the absence of other information, it would appear that it is best to precondition the matrix A so that all the diagonal elements are equal, e.g. the covariance matrix should be replaced by the correlation matrix.

The problem of preconditioning symmetric positive definite matrices arise in the other statistical contexts (cf. [12]).

5. Iterative refinement

Once an approximate solution to $A\underset{\sim}{x} = \underset{\sim}{b}$ has been obtained, it is frequently possible to improve the accuracy of the approximate solution. Let $\bar{\underset{\sim}{x}}$ be an approximate solution, and let $\underset{\sim}{r} = \underset{\sim}{b} - A\bar{\underset{\sim}{x}}$. Then if $\bar{\underset{\sim}{x}} = \underset{\sim}{x} + \underset{\sim}{\delta}$, $\underset{\sim}{\delta}$ satisfies the equation

$$A\underset{\sim}{\delta} = \underset{\sim}{r} . \tag{5.1}$$

Equation (5.1) can be solved approximately once the Cholesky decomposition of A is known; indeed, it requires but $n^2 + O(n)$ multiplications to solve for the correction $\underset{\sim}{\delta}$. Of course, it is not possible to solve precisely for $\underset{\sim}{\delta}$ so that the process may be repeated. Thus for $\underset{\sim}{x}^{(0)}$ given, the algorithm proceeds as follows:

1) compute $\underset{\sim}{r}^{(k)} = \underset{\sim}{b} - A\underset{\sim}{x}^{(k)}$;

2) solve $A\underset{\sim}{\delta}^{(k)} = \underset{\sim}{r}^{(k)}$;

3) compute $\underset{\sim}{x}^{(k+1)} = \underset{\sim}{x}^{(k)} + \underset{\sim}{\delta}^{(k)}$.

The process continues until

$$\frac{\|\underset{\sim}{x}^{(k+1)} - \underset{\sim}{x}^{(k)}\|}{\|\underset{\sim}{x}^{(k+1)}\|} < \varepsilon ,$$

a predetermined constant or some other criterion is satisfied. The above algorithm is known as iterative refinement and has been extensively discussed in the literature (cf. [24, 32]) .

There are three sources of error in the process: (1) computation of the residual vector $\underset{\sim}{r}^{(k)}$, (2) solution of the system of equations

for the correction vector $\underset{\sim}{\delta}^{(k)}$, and (3) addition of the correction vector to the approximation $\underset{\sim}{x}^{(k)}$. It is absolutely necessary to compute the. components of the residual vector using double precision inner products and then to round to single precision accuracy. The convergence of the iterative refinement process has been discussed in detail by Moler [24]. Generally speaking, for a large class of matrices for $k \geq k_0$ all components of $\underset{\sim}{x}^{(k)}$ are the correctly rounded single precision approximations to the components of $\underset{\sim}{x}$. There are exceptions to this, however, (cf.[21]). Experimentally, it has been observed, in most instances, that if $\|\underset{\sim}{\delta}^{(0)}\|_\infty / \|\underset{\sim}{x}^{(0)}\|_\infty \leq 2^{-p}$ where

$$\|\underset{\sim}{y}\|_\infty = \max_{1 \leq i \leq n} |y_i| ,$$

then $k_0 \geq [t/p]$. We shall return to the subject of iterative refinement when we discuss the solution of linear least squares problem.

6. Partial Correlation

Again let A be a positive definite matrix and we partition the matrix in the following form:

$$A = \left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right]$$

where A_{11} is p x p , A_{22} is q x q , and $A_{12}^{T} = A_{21}$. Suppose the SCDA is used but the algorithm is stopped after p steps, Then

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} R_1^T \\ \cdots \\ S^T \end{bmatrix} \left[\begin{array}{c:c} R_1 & S \end{array}\right] + \begin{bmatrix} 0 & 0 \\ 0 & W \end{bmatrix} ;$$

here R_1 is the Cholesky factor of A_{11} so that $R_1^T R_1 = A_{11}$. Equating matrix blocks, we see

$$A_{12} = R_1^T S$$

$$A_{22} = S^T S + W .$$

Thus

$$W = A_{22} - A_{21} R_1^{-1} R_1^{-T} A_{12}$$

$$= A_{22} - A_{21} A_{11}^{-1} A_{12} .$$

The matrix W is denoted by $a_{ij}^{(p+1)}$ in (1.2).

Consider the covariance matrix

$$\Sigma = \begin{pmatrix} \Sigma_{11} \Sigma_{12} \\ \Sigma_{21} \Sigma_{22} \end{pmatrix} \text{ where } \Sigma_{11} \text{ is } p \times p .$$

The partial covariance matrix

$$\Sigma_{22 \cdot 1} = \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}$$

when the first p variables are held fixed, and the regression function is defined by

$$\mu^{(2)} + \Sigma_{21} \Sigma_{11}^{-1} (\underset{\sim}{x}^{(1)} - \mu^{(1)})$$

where $\mu^{(1)}$, $\mu^{(2)}$ are the corresponding vectors of expected values. Thus if we apply the first p steps of the SCDA, $\sigma_{ij}^{(p+1)}$ corresponds to the partial covariance when the first p variables are eliminated. We can eliminate the effect of the first (p+1) variables by simply performing one more step of the SCDA. It is a simple matter to compute the regression function since $\Sigma_{21} \Sigma_{11}^{-1}$ corresponds to $S^T R_{\sim}^{-1}$.

7. Least squares

Let A be a given $m \times n$ real matrix of rank r and $\underset{\sim}{b}$ a given vector. We wish to determine $\hat{\underset{\sim}{x}}$ such that

$$\sum_{i=1}^{m} (b_i - \sum_{i=1}^{n} a_{ij}x_j)^2 = \min .$$

or using matrix notation

$$\|\underset{\sim}{b}-A\underset{\sim}{x}\|_2 = \min . \qquad (7.1)$$

If $m \geq n$ and $r < n$, then there is no unique solution. Under these conditions, we require amongst those vectors $\underset{\sim}{x}$ which satisfy (7.1) that

$$\|\underset{\sim}{x}\|_2 = \min .$$

For $r=n$, $\hat{\underset{\sim}{x}}$ satisfies the normal equations

$$A^TA\hat{\underset{\sim}{x}} = A^Tb . \qquad (7.2)$$

Unfortunately, the matrix A^TA is frequently ill-conditioned and influenced greatly by roundoff errors. The following example illustrates this well. Suppose

$$A = \begin{vmatrix} 1 & 1 & 1 & 1 \\ \varepsilon & 0 & 0 & 0 \\ 0 & \varepsilon & 0 & 0 \\ 0 & 0 & \varepsilon & 0 \\ 0 & 0 & 0 & \varepsilon \end{vmatrix}$$

which is clearly of rank 4. Then

$$A^TA = \begin{bmatrix} 1+\varepsilon^2 & 1 & 1 & 1 \\ 1 & 1+\varepsilon^2 & 1 & 1 \\ 1 & 1 & 1+\varepsilon^2 & 1 \\ 1 & 1 & 1 & 1+\varepsilon^2 \end{bmatrix}$$

and the eigenvalues of A^TA are $4+\varepsilon^2, \varepsilon^2, \varepsilon^2, \varepsilon^2$. Assume that the elements of A^TA are computed using double precision arithmetic, and then rounded to single precision accuracy. Now if $\varepsilon < \sqrt{2^{-t}}$,

$$fl(A^TA) = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix},$$

a matrix of rank one, and consequently, no matter how accurate the linear equation solver it will be impossible to solve the normal equations (7.2).

Longley [23] has given examples in which the solution of the normal equations leads to almost no digits of accuracy of the least squares problem.

8. A matrix decomposition

Now $\|\underset{\sim}{y}\|_2 = (\underset{\sim}{y}^T\underset{\sim}{y})^{1/2}$ so that $\|Q\underset{\sim}{y}\|_2 = \|\underset{\sim}{y}\|_2$ when Q is an orthogonal matrix, viz, $Q^TQ = I$. Thus

$$\|\underset{\sim}{b}-A\underset{\sim}{x}\|_2 = \|\underset{\sim}{c}-QA\underset{\sim}{x}\|_2$$

where $\underset{\sim}{c} = Q\underset{\sim}{b}$ and Q is an orthogonal matrix. We choose Q so that

$$QA = R = \begin{pmatrix} \tilde{R} \\ \cdots \\ 0 \end{pmatrix} \Big\}(m-n) \times n \qquad (8.1)$$

where $\tilde{R}$ is an upper triangular matrix. Let

$$\tilde{R} = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ & r_{22} & \cdots & r_{2n} \\ & \bigcirc & \ddots & \\ & & & r_{nn} \end{pmatrix},$$

then

$$\|\underset{\sim}{b}-A\underset{\sim}{x}\|_2^2 = (c_1-r_{11}x_1-r_{12}x_2 - \ldots - r_{1n}x_n)^2$$
$$+ (c_2-r_{22}x_2 - \ldots - r_{2n}x_n)^2$$
$$+ \ldots + (c_n-r_{nn}x_n)^2$$
$$+ c_{n+1}^2 + c_{n+2}^2 + \ldots + c_m^2 .$$

Thus $\|\underset{\sim}{b}-A\underset{\sim}{x}\|_2^2$ is minimized when

$$
\begin{aligned}
r_{11}\hat{x}_1 + r_{12}x_2 + \cdots + r_{1n}\hat{x}_n &= c_1 \\
r_{22}x_2 + \cdots + r_{2n}x_n &= c_2 \\
&\vdots \\
r_{nn}x_n &= c_n
\end{aligned}
$$

i.e., $\tilde{R}\hat{\underset{\sim}{x}} = \underset{\sim}{\tilde{c}}$ where

$$\underset{\sim}{\tilde{c}}^T = (c_1, c_2, \ldots, c_n) ,$$

and

$$\|\underset{\sim}{b}-A\hat{\underset{\sim}{x}}\|_2^2 = c_{n+1}^2 + c_{n+2}^2 + \cdots + c_m^2 . \qquad (8.2)$$

Then

$$
\begin{aligned}
R^TR = (\tilde{R}\!:\!0)^T(\tilde{R}\!:\!0) &= \tilde{R}^T\tilde{R} \\
= (QA)^T(QA) &= A^TA ,
\end{aligned}
\qquad (8.3)
$$

and thus $\tilde{R}^T\tilde{R}$ is simply the Cholesky decomposition of A^TA .

There are a number of ways to achieve the decomposition (8.1); e.g., one could apply a sequence of plane rotations to annihilate the elements below the diagonal of A . A very effective method to realize the decomposition (8.1) is via Householder transformations. A matrix P is said to be a Householder transformation if

$$P = I - 2\underset{\sim}{u}\underset{\sim}{u}^T , \quad \underset{\sim}{u}^T\underset{\sim}{u} = 1 .$$

Note that 1) $P = P^T$ and 2) $PP^T = I - 2\underset{\sim}{u}\underset{\sim}{u}^T - 2uu^T + 4\underset{\sim}{u}\underset{\sim}{u}^T\underset{\sim}{u}\underset{\sim}{u}^T = I$ so that P is a symmetric, orthogonal transformation.

Let $A^{(1)} = A$ and let $A^{(2)}, A^{(3)}, \ldots A^{(n+1)}$ be defined as follows:

$$A^{(k+1)} = P^{(k)}A^{(k)} \qquad (k=1,2,\ldots,n)$$

where $P^{(k)} = I - 2\underset{\sim}{w}^{(k)}\underset{\sim}{w}^{(k)T}$, $\underset{\sim}{w}^{(k)T}\underset{\sim}{w}^{(k)} = 1$. The matrix $P^{(k)}$ is chosen so that $a_{k+1,k}^{(k+1)} = a_{k+2,k}^{(k+1)} = \ldots = a_{n,k}^{(k+1)} = 0$. Thus after k transformations

$$A^{(k+1)} = \begin{vmatrix} a_{11}^{(2)} & a_{12}^{(2)} & \cdots & \cdots & a_{1n}^{(2)} \\ 0 & a_{22}^{(3)} & \cdots & \cdots & a_{2n}^{(3)} \\ 0 & \ddots & \ddots & \ddots & \\ \vdots & & a_{kk}^{(k+1)} & \cdots & a_{kn}^{(k+1)} \\ \vdots & & 0 & a_{k+1,k+1}^{(k+1)} & \cdots \\ \vdots & & 0 & \vdots & \vdots \\ 0 & & 0 & a_{m,k+1}^{(k+1)} & a_{mn}^{(k+1)} \end{vmatrix}$$

The details of the computation are given in [5] and [13].

Clearly,

$$R = A^{(n+1)}$$

and

$$Q = P^{(n)}P^{(n-1)} \cdots P^{(1)}$$

although one need not compute Q explicitly. The number of multiplications required to produce R is roughly $mn^2 - \frac{n^3}{3}$ whereas approximately $\frac{mn^2}{2}$ multiplications are required to form the normal equations (7.2).

9. Statistical calculations

In many statistical calculations, it is necessary to compute certain auxiliary information associated with A^TA. These can readily be obtained from the orthogonal decomposition. Thus

$$\det(A^TA) = (r_{11} x r_{22} x \ldots x r_{nn})^2.$$

Since

$$A^TA = \tilde{R}^T\tilde{R} \ , \ (A^TA)^{-1} = \tilde{R}^{-1}\tilde{R}^{-T} \ .$$

The inverse of $\tilde{R}$ can be readily obtained since $\tilde{R}$ is an upper triangular matrix. It is possible to calculate $(A^TA)^{-1}$ directly from $\tilde{R}$. Let

$$(A^TA)^{-1} = x = (\underset{\sim}{x}_1, \underset{\sim}{x}_2, \ldots, \underset{\sim}{x}_n).$$

Then from the relationship

$$\tilde{R} X = \tilde{R}^{-T}$$

and by noting that $\{\tilde{R}^{-T}\}_{ii} = 1/r_{ii}$, it is possible to compute $\underset{\sim}{x}_n, \underset{\sim}{x}_{n-1}, \ldots, \underset{\sim}{x}_1$. The number of operations are roughly the same as in the first method but more accurate bounds may be established for this method provided all inner products are accumulated to double precision.

In some applications, the original set of observations are augmented by an additional set of observations. In this case, it is not necessary to begin the calculation from the beginning again if the method of orthogonalization is used. Let $\tilde{R}_1, \underset{\sim}{c}_1$ correspond to the original data after it has been reduced by orthogonal transformations and let $A_2, \underset{\sim}{b}_2$ correspond to the additional observations. Then the up-dated least squares solution can be obtained directly from

$$A = \begin{pmatrix} A_2 \\ \cdots \\ \tilde{R}_1 \end{pmatrix}, \qquad b = \begin{pmatrix} b_2 \\ \cdots \\ c_1 \end{pmatrix}$$

This follows immediately from the fact that the product of two orthogonal transformations is an orthogonal transformation.

The above observation has another implication. One of the arguments frequently advanced for using normal equations is that only $n(n+1)/2$ memory locations are required. By partitioning the matrix A by rows, however, then similarly only $n(n+1)/2$ locations are needed when the method of orthogonalization is used.

In certain statistical applications, it is desirable to remove a row of the matrix A after the least squares solution has been obtained. This can be done in a very simple manner. Consider the matrix

$$A = \begin{pmatrix} \tilde{R} \\ \cdots\cdots \\ i\ \underset{\sim}{\alpha} \end{pmatrix} \qquad \text{and } \underset{\sim}{d} = \begin{pmatrix} \tilde{c} \\ \cdots \\ i\ \beta \end{pmatrix} .$$

where $\underset{\sim}{\alpha}$ is the row of A which one wishes to remove, β is the corresponding element of $\underset{\sim}{b}$, and $i = \sqrt{-1}$. Note that

$$S^T S = \tilde{R}^T \tilde{R} - \underset{\sim}{\alpha}^T \underset{\sim}{\alpha} = A^T A - \underset{\sim}{\alpha}^T \underset{\sim}{\alpha}$$

Let

$$z_{1,n+1} = \begin{bmatrix} r\cos\theta & & & \sin\theta \\ & 1 & & \\ & & \ddots & \\ & & & 1 \\ \sin\theta & & & -\cos\theta \end{bmatrix}$$

$$S^{(1)} = S \quad , \text{ and } S^{(2)} = Z_{1,n+1} S^{(1)} .$$

We choose cos θ so that $\{S^{(2)}\}_{n+1,1} = 0$. Thus

$$\{S^{(2)}\}_{1,1} = \sqrt{r_{11}^2 - \alpha_1^2}$$

$$\{S^{(2)}\}_{1,j} = \frac{r_{11}r_{1j} - \alpha_1\alpha_j}{\sqrt{r_{11}^2 - \alpha_1^2}} \qquad j = 2,3,\ldots,n$$

$$\{S^{(2)}\}_{n+1,j} \quad \frac{i(\alpha_1 r_{1j} - \alpha_j r_{11}}{\sqrt{r_{11}^2 - \alpha_1^2}} \qquad j = 2,3,\ldots,n \ .$$

Note no complex arithmetic is really necessary.

The process is continued as follows:

Let

$$Z_{k,n+1} = \begin{bmatrix} & & k & & n+1 & \\ 1 & & & & & \\ & 1 & & & & \\ & & \cos\theta_k & & \sin\theta_k & k \\ & & & 1 & & \\ & & & & 1 & \\ & & \sin\theta_k & & -\cos\theta_k & n+1 \end{bmatrix}$$

Then

$$S^{(k+1)} = Z_{k,n+1}S^{(k)} \qquad , \qquad k = 1,2,\ldots,n \ ,$$

and cos θ_k is determined so that $\{S^{(k+1)}\}_{k,n+1} = 0$. Thus roughly $3n^2$ multiplications and divisions and n square roots are required to form the new $\tilde{R}$.

Suppose it is desirable to add an additional variable so that the matrix A is augmented by a vector $\underset{\sim}{g}$ (say). The first n columns of $\tilde{R}_n$ are unchanged. Now one computes

$$\underset{\sim}{h} = P^{(\quad)} \ldots P^{(2)}P^{(1)}\ \underset{\sim}{g} .$$

From $\underset{\sim}{h}$ one can compute $P^{(n+1)}$ and apply it to $P^{(n)} \ldots P^{(1)}\ \underset{\sim}{b}$.

It is also possible to drop one of the variables in a simple fashion after $\tilde{R}$ has been computed. For example, suppose we wish to drop variable 1, then

$$\tilde{R} = \begin{bmatrix} r_{12} & \cdot & r_{1n} \\ r_{22} & \cdot & \\ & \cdot & \\ 0 & \cdot & \\ & & r_{nn} \end{bmatrix}_{n \times (n-1)}$$

By using plane rotations, similar to those given by (9.1), it is possible to reduce $\tilde{R}$ to the triangular form again.

10. Gram-Schmidt orthogonalization

In section 8, it was shown that it is possible to write

$$QA = R \,. \tag{10.1}$$

The matrix Q is constructed as a product of Householder transformation. From (10.1), we see that

$$A = Q^T R \equiv PS$$

where $P^T P = I_n$, $S: \triangle$. Each row of S and each column of P is uniquely determined up to a scalar factor of modulus one. In order to avoid computing square roots, we modify the algorithms so that S is an upper traingular matrix with ones on the diagonal. Thus $P^T P = D$, a diagonal matrix. The calculation of P and S may be calculated in two ways.

a) Classical Gram-Schmidt Algorithm (CGSA)

The elements of S are computed one column at a time. Let

$$A^{(k)} = (\underset{\sim}{p}_1, \underset{\sim}{p}_2, \cdots, \underset{\sim}{p}_{k-1}, \underset{\sim}{a}_k, \cdots, \underset{\sim}{a}_n)$$

and assume

$$\underset{\sim}{p}_i^T \underset{\sim}{p}_j = \delta_{ij} d_i \,, \quad 1 \leq i,j < k-1.$$

At step k, we compute

$$s_{ik} = \underset{\sim}{p}_i^T \underset{\sim}{a}_k / d_i \,, \qquad 1 \leq i \leq k-1$$

$$\underset{\sim}{p}_k = \underset{\sim}{a}_k - \sum_{i=1}^{k-1} s_{ik} \underset{\sim}{p}_i \,, \; d_k = \|\underset{\sim}{p}_k\|_2^2 \,.$$

b) Modified Gram-Schmidt Algorithm (MGSA)

Here the elements of S are computed one row at a time. We define

$$A^{(k)} = (\underset{\sim}{p}_1, \underset{\sim}{p}_2, \cdots, \underset{\sim}{p}_{k-1}, \underset{\sim}{a}_k^{(k)}, \cdots, \underset{\sim}{a}_n^{(k)})$$

and assume

$$\underset{\sim}{p}_i^T \underset{\sim}{p}_j = \delta_{ij} d_i, \quad \underset{\sim}{q}_i^T \underset{\sim}{a}_\ell^{(k)} = 0 \quad , 1 \leq i, j \leq k-1, \; k < \ell \leq n.$$

At step k, we take $\underset{\sim}{p}_k = \underset{\sim}{a}_k^{(k)}$, and compute

$$d_k = \|\underset{\sim}{p}_k\|_2^2 \quad , \quad s_{k\ell} = \underset{\sim}{p}_k^T \underset{\sim}{a}_\ell^{(k)} / d_k \quad , \quad \underset{\sim}{a}_\ell^{(k+1)} = \underset{\sim}{a}_\ell^{(k)} - s_{k\ell} \underset{\sim}{p}_k \quad ,$$

$$k+1 \leq \ell \leq n.$$

In both procedures, $s_{kk} = 1$. The two procedures in the absence of roundoff errors, produce the same decomposition. However, they have completely different numerical properties when $n > 2$. If A is at all "ill-conditioned", then using the CGSA, the computed columns of P will soon lose their orthogonality. Consequently, one should never use the CGSA without reorthogonalization, which greatly increases the amount of computation. Reorthogonalization is never needed when using the MGSA. A careful roundoff analysis is given by Björck in [2]. Rice [27] has shown experimentally that the MGSA produces excellent results.

The MGSA has the advantages that it is relatively easy to program, and experimentally (cf. [20]), it seems to be slightly more accurate than the Householder procedure. However, it requires roughly $\frac{mn^2}{2}$ operations which is slightly more than that necessary to the Householder procedure. Furthermore, it is not as simple as the Householder procedure to add observations.

11. Sensitivity of the solution

We consider first the inherent sensitivity of the solution of the least squaresproblem. For this purpose it is convenient to introduce the condition number K(A) of a non-square matrix A . This is defined by

$$K(A) = \sigma_1/\sigma_n \,,\quad \sigma_1 = \max_{x\neq 0}\|A\,\underset{\sim}{x}\|_2/\|\,\underset{\sim}{x}\,\|_2 \,,\quad \sigma_n = \min_{x\neq 0}\|A\,\underset{\sim}{x}\|_2/\|\,\underset{\sim}{x}\,\|_2$$

so that σ_1^2 and σ_n^2 are the greatest and the least eigenvalues of $A^T A$. From its definition it is clear that K(A) is invariant with respect to unitary transformations. If $\tilde{R}$ is defined as in (8.1) then

$$\sigma_1(\tilde{R}) = \sigma_1(A) \,,\quad \sigma_n(\tilde{R}) = \sigma_n(A) \,,\quad \kappa(\tilde{R}) = \kappa(A),$$

while

$$\sigma_1(\tilde{R}) = \|\,\tilde{R}\,\|_2 \text{ and } \sigma_n(\tilde{R}) = 1/\|\,\tilde{R}^{-1}\,\|_2 \,.$$

The commonest method of solving least squares problems is via the normal equation

$$A^T A\,\underset{\sim}{x} = A^T\,\underset{\sim}{b} \quad . \tag{11.1}$$

The matrix $A^T A$ is square and we-have

$$\kappa(A^T A) = \kappa^2(A) \,.$$

-This means that if A has a condition number of the order of $2^{\frac{1}{2}t}$ then $A^T A$ has a condition number of order 2^t and it will not be possible using t -digit arithmetic to solve (11.1). The method of orthogonal transformations replaces the least squares problem by the solution of

the equations $\tilde{R}\,\underset{\sim}{x} = \underset{\sim}{\tilde{c}}$ and $\kappa(\tilde{R}) = \kappa(A)$. It would therefore seem to have substantial advantages since we avoid working with a matrix with condition number $\kappa^2(A)$.

We now show that this last remark is an oversimplification. To this end, we compare the solution of the original system $(A \vdots \underset{\sim}{b})$ with that of a perturbed system. It is convenient to assume that

$$\sigma_1 = \|A\|_2 = \| \underset{\sim}{b} \|_2 = 1 ;$$

this is not in any sense a restriction since we can make $\|A\|_2$ and $\| \underset{\sim}{b} \|_2$ of order unity merely by scaling by an appropriate power of two. We now have

$$\kappa(A) = \kappa(\tilde{R}) = \| \tilde{R}^{-1} \|_2 = 1/\sigma_n .$$

Consider the perturbed system

$$(A + \varepsilon E \vdots \underset{\sim}{b} + \varepsilon \underset{\sim}{e}) , \|E\|_2 = \| \underset{\sim}{e} \|_2 = 1 ,$$

where ε is to be arbitrarily small. The solution $\underset{\sim}{\bar{x}}$ of the perturbed system satisfies the equation

$$(A + \varepsilon E)^T (A + \varepsilon E)\, \underset{\sim}{\bar{x}} = (A + \varepsilon E)^T (\underset{\sim}{b} + \varepsilon \underset{\sim}{e}) . \qquad (11.2)$$

If $\underset{\sim}{\hat{x}}$ is the exact solution of the original system and Q is the exact orthogonal transformation corresponding to A we have

$$Q\,A = \begin{bmatrix} \tilde{R} \\ \cdots \\ 0 \end{bmatrix} , \quad Q(A + \varepsilon E) = \begin{bmatrix} \tilde{R} + \varepsilon F \\ \cdots\cdots \\ \varepsilon G \end{bmatrix} , \quad Q\,\underset{\sim}{e} = \begin{bmatrix} \underset{\sim}{f} \\ \vdots \\ \underset{\sim}{g} \end{bmatrix}$$

and

$$\underset{\sim}{r} = \underset{\sim}{b} - A\,\hat{\underset{\sim}{x}} \quad , \quad A^T\underset{\sim}{r} = \underset{\sim}{\theta} .$$

Equation (11.2) therefore becomes

$$(A + \varepsilon E)^T(A + \varepsilon E) = (A^T + \varepsilon E^T)(A\,\underset{\sim}{x} + \underset{\sim}{r} + s\underset{\sim}{e})$$

giving

$$\begin{bmatrix} \tilde{R} + \varepsilon F \\ \cdots\cdots \\ \varepsilon G \end{bmatrix}^T \begin{bmatrix} \tilde{R} + \varepsilon F \\ \cdots\cdots \\ \varepsilon G \end{bmatrix} \bar{\underset{\sim}{x}} = \begin{bmatrix} \tilde{R} + \varepsilon F \\ \cdots\cdots \\ \varepsilon G \end{bmatrix}^T \left(\begin{bmatrix} \tilde{R} \\ \cdots \\ 0 \end{bmatrix} \underset{\sim}{x} + \varepsilon \begin{bmatrix} \underset{\sim}{f} \\ \cdots \\ \underset{\sim}{g} \end{bmatrix} \right) + \varepsilon E^T \underset{\sim}{r} .$$

Neglecting ε^2 where advantageous

$$(\tilde{R} + \varepsilon F)^T(\tilde{R} + \varepsilon F)\,\bar{\underset{\sim}{x}} = (\tilde{R} + \varepsilon F)^T\tilde{R}\,\hat{\underset{\sim}{x}} + \varepsilon(\tilde{R} + \varepsilon F)^T\underset{\sim}{f} + \varepsilon E^T\underset{\sim}{r} + O(\varepsilon^2)$$

$$\underset{\sim}{x} = (\tilde{R} + \varepsilon F)^{-1}\,\tilde{R}\,\hat{\underset{\sim}{x}} + \varepsilon(\tilde{R} + \varepsilon F)^{-1}\underset{\sim}{f} +$$

$$+ \varepsilon(\tilde{R}^T\tilde{R})^{-1}E^T\underset{\sim}{r} + O(\varepsilon^2)$$

$$= \hat{\underset{\sim}{x}} - \varepsilon\tilde{R}^{-1}F\,\hat{x} + \varepsilon R^{-1}\underset{\sim}{f} + \varepsilon(\tilde{R}^T\tilde{R})^{-1}E^T\underset{\sim}{r} + O(\varepsilon^2)$$

giving

$$\| \bar{\underset{\sim}{x}} - \hat{\underset{\sim}{x}} \|_2 \le \varepsilon\|\tilde{R}^{-1}\|_2\|F\|_2\| \hat{\underset{\sim}{x}} \|_2 + \varepsilon\|\tilde{R}^{-1}\|_2\|\underset{\sim}{f}\|_2 + \varepsilon\|\tilde{R}^{-1}\|_2^2\,\|E\|_2\|\underset{\sim}{r}\|_2 + O(\varepsilon^2)$$

$$\le \varepsilon\kappa(A)\|\hat{\underset{\sim}{x}}\|_2 + \varepsilon\kappa(A) + \varepsilon\kappa^2(A)\|\underset{\sim}{r}\|_2 + O(\varepsilon^2) \quad .$$

We observe that the bounds include a term $\varepsilon\kappa^2(A)\|\underset{\sim}{r}\|_2$. It is easy to verify by means of a 3 X 2 matrix A that this bound is realistic and that an error of this order of magnitude does indeed result from almost any such perturbation **E of A** . We conclude that although the use of the orthogonal transformation avoids some of the ill effects inherent in

the use of the normal equations the value of $\kappa^2(A)$ is still relevant to some extent.

When the equations are compatible $\|\underset{\sim}{r}\| = 0$ and the term in $\kappa^2(A)$ disappears. In the non-singular linear equation case $\underset{\sim}{r}$ is always null and hence it is always $\kappa(A)$ rather than $\kappa^2(A)$ which is relevant. Since the sensity of the solution depends on the condition number, in the absence of other information, one should normalize each column of A so that its length is one in accordance with (4.1).

12. Iterative refinement for least squares problems

The iterative refinement method may also be used for improving the solution to linear least squares problems.

Let

$$\alpha \underset{\sim}{\rho} = \underset{\sim}{b} - A \underset{\sim}{\hat{x}} \qquad , \quad \alpha > 0$$

so that

$$\alpha A^T \rho = A^T b - A^T A \underset{\sim}{\hat{x}} = \underset{\sim}{\theta} \ .$$

When $\alpha = 1$, the vector $\underset{\sim}{\rho}$ is simply the residual vector $\underset{\sim}{r}$. Thus

$$\left[\begin{array}{c|c} \alpha I & A \\ \hline A^T & 0 \end{array}\right] \left[\begin{array}{c} \underset{\sim}{\rho} \\ \hline \underset{\sim}{\hat{x}} \end{array}\right] = \left[\begin{array}{c} \underset{\sim}{b} \\ \hline \underset{\sim}{\theta} \end{array}\right] \ , \qquad (12.1)$$

or

$$B \underset{\sim}{y} = \underset{\sim}{g} \ .$$

One of the standard methods for solving linear equations may now be used to solve (12.1). However, this is quite wasteful of memory space since the dimension of the system to be solved is (m + n).

We may simplify this problem somewhat by noting with the aid of (8.3) that

$$\left[\begin{array}{c|c} \alpha I & A \\ \hline A^T & 0 \end{array}\right] = \left[\begin{array}{c|c} \sqrt{\alpha}\, I & 0 \\ \hline \frac{1}{\sqrt{\alpha}} A^T & \frac{1}{\sqrt{\alpha}} \tilde{R}^T \end{array}\right] \left[\begin{array}{c|c} \sqrt{\alpha}\, I & \frac{1}{\sqrt{\alpha}} A \\ \hline 0 & -\frac{1}{\sqrt{\alpha}} \tilde{R} \end{array}\right] \equiv LU \ .$$

We are now in a position to use the iterative refinement method for solving linear equations.

Thus one might proceed as follows:

1) Solve for $\underset{\sim}{x}^{(o)}$ using one of the orthogonalization procedures outlined in section 8 or 10. $\mathfrak{R}$ must be saved but it is not necessary to retain Q. Then

$$\underset{\sim}{\rho}^{(o)} = \frac{1}{\alpha}(\underset{\sim}{b} - A\,\underset{\sim}{x}^{(o)}) \ .$$

2) The vector $\underset{\sim}{y}^{(s+1)}$ is determined from the relationship

$$\underset{\sim}{y}^{(s+1)} = \underset{\sim}{y}^{(s)} + \underset{\sim}{\delta}^{(s)}$$

where

$$B\,\underset{\sim}{\delta}^{(s)} = \underset{\sim}{g} - B\,\underset{\sim}{y}^{(s)} = \underset{\sim}{h}^{(s)} \quad . \tag{12.2}$$

This calculation is simplified by solving

$$L\,\underset{\sim}{z}^{(s)} = \underset{\sim}{h}^{(s)}$$

$$U\,\underset{\sim}{\delta}^{(s)} = \underset{\sim}{z}^{(s)} \quad .$$

The vector $\underset{\sim}{h}^{(s)}$ must be calculated using double precision accuracy and then rounding to single precision.

3) Terminate the iteration when $\|\underset{\sim}{\delta}^{(s)}\|/\|\underset{\sim}{y}^{(s)}\|$ is less than a prescribed number.

Note that the computed residual vector is an approximation to the residual vector when the exact solution $\hat{\underset{\sim}{x}}$ is known. This may differ from the residual vector computed from the approximate solution to the

least squares problem.

A variant of the above procedure has been analyzed by Björck [3], and he has also given an ALGOL procedure. This has proved to be a very effective method for obtaining highly accurate solutions to linear least squares problems. Björck and Golub [4] have described a similar iterative refinement method for solving least squares problems with linear constraints.

13. Singular Systems

If the rank of A is less than n and if column interchanges are performed to maximize the diagonal elements of R, then

$$A^{(r+1)} = \left[\begin{array}{c|c} \tilde{R}_{rxr} & S_{(n-r)xr} \\ \hline 0 & 0 \end{array}\right]^{1}$$

when rank (A) = r. A sequence of Householder transformations may now be applied on the right of $A^{(r+1)}$ so that the elements of $S_{(n-r)Xr}$ become annihilated. Thus dropping subscripts and superscripts, we have

$$Q\ A\ Z = T = \left[\begin{array}{c|c} \tilde{T} & 0 \\ \hline 0 & 0 \end{array}\right]$$

where $\tilde{T}$ is an rxr upper triangular matrix. Now

$$\|\underset{\sim}{b} - A\ \underset{\sim}{x}\|_2 = \|\underset{\sim}{b} - Q^T\ T\ Z^T \underset{\sim}{x}\|_2$$

$$= \|\underset{\sim}{c} - T\ \underset{\sim}{y}\|_2$$

where $\underset{\sim}{c} = Q\underset{\sim}{b}$ and $\underset{\sim}{y} = Z^T\underset{\sim}{x}$. Since T is of rank r, there is no unique solution so that we impose the condition that $\|\ \hat{\underset{\sim}{x}}\ \|_2$ = min. But $\|\ \underset{\sim}{y}\ \|_2 = \|x\|_2$ since T is orthogonal and $\|\ \underset{\sim}{y}\ \|_2$ = min. when

$$y_{r+1} = y_{r+2} = \ldots = y_m = 0.$$

Thus

$$\hat{\underset{\sim}{x}} = Z\left(\begin{array}{c|c} \tilde{T}^{-1} & 0 \\ \hline 0 & 0 \end{array}\right) Q\ \underset{\sim}{b}\ .$$

This solution has been given by Fadeev, *et. al.* [7] and Hanson and Lawson [18]. The problem still remains how to numerically determine the rank which will be discussed in section 15.

14. Singular value decomposition

Let A be a real, m x n matrix (for notational convenience we assume that $m \geqq n$). It is well known (cf. [22]) that

$$A = U\Sigma V^T \tag{14.1}$$

where

$$UU^T = I_m \ , \ VV^T = I_n$$

and

$$\Sigma = \begin{pmatrix} \sigma_1, \cdots, 0 \\ \cdots\cdots\cdots \\ 0, \ \cdots, \sigma_n \\ \hline 0 \end{pmatrix} \quad \} \ (m - n) \times n \ .$$

The matrix U consists of the orthonormalized eigenvectors of AA^T, and the matrix V consists of the orthonormalized eigenvectors of A^TA. The diagonal elements of Σ are the non-negative square roots of the eigenvalues of A^TA; they are called singular values or principal values of A. We assume

$$\sigma_1 \geqq \sigma_2 \geqq \cdots \geqq \sigma_n \geqq 0.$$

Thus if rank (A) = r, $\sigma_{r+1} = \sigma_{r+2} = \cdots = \sigma_n = 0$. The decomposition (14.1) is called the singular value decomposition (SVD).

Let

$$\tilde{A} = \begin{pmatrix} 0 & A \\ A^T & 0 \end{pmatrix} \ . \tag{14.2}$$

It can be shown [22] that the non-zero eigenvalues of $\tilde{A}$ always occur in + pairs, viz

$$\lambda_j(\tilde{A}) = \pm \sigma_j(A) \quad (j = 1,2,\ldots,r). \tag{14.3}$$

15. Applications of the SVD

The singular value decomposition plays an important role in a number of least squares problems, and we will illustrate this with some examples. Throughout this discussion, we use the Euclidean or Frobenius norm of a matrix, viz.

$$\|A\| = (\sum |a_{ij}|^2)^2$$

A) Let $\mathcal{U}_n$ be the set of all n x n orthogonal matrices. For an arbitrary n x n real matrix A, determine $Q \varepsilon \mathcal{U}_n$ such that

$$\|A - Q\| \leqq \|A - X\| \text{ for any } X \varepsilon \mathcal{U}_n .$$

It has been shown by Fan and Hoffman [8] that if

$$A = U\Sigma V^T \text{ , then } Q = UV^T .$$

B) An important generalization of problem A occurs in factor analysis. For arbitrary n X n real matrices A and B, determine $Q \varepsilon \mathcal{U}_n$ such that

$$\|A - BQ\| \leqq \|A - BX\| \text{ for any } X \varepsilon \mathcal{U}_n .$$

It has been shown by Green [17] and by Schönemann [28] that if

$$B^T A = U\Sigma V^T, \text{ then } Q = UV^T .$$

C) Let $\mathcal{M}_{m,n}^{(k)}$ be the set of all m x n matrices of rank k. Assume $A \mathcal{M}_{m,n}^{(r)}$. Determine $B \varepsilon \mathcal{M}_{m,n}^{(k)}$ $(k \leqq r)$ such that

$$\|A - B\| \leqq \|A - X\| \text{ for all } X \varepsilon \mathcal{M}_{m,n}^{(k)} .$$

It has been shown by Eckart and Young [6] that if

$$A = U\Sigma V^T, \text{ then } B = U\Omega_k V^T \tag{15.1}$$

where

$$\Omega_k = \begin{pmatrix} \sigma_1 & & 0 \\ & \sigma_2 & \\ & \ddots & \\ 0 & & \sigma_k \\ \hline & & \\ & 0 & \end{pmatrix}. \tag{15.2}$$

Note that

$$\|A - B\| = \|\Sigma - \Omega_k\| = (\sigma_{k+1}^2 + \ldots + \sigma_r^2)^{\frac{1}{2}} \quad . \tag{15.3}$$

D) An n x m matrix X is said to be the psuedo-inverse of an m X n matrix A if X satisfies the following four properties:

i) $AXA = A$,

ii) $XAX = X$,

iii) $(AX)^T = Ax$,

iv) $(XA)^T = XA$.

We denote the psuedo-inverse by A^+ . We wish to determine A^+ numerically. It can be shown [26] that A^+ can always be determined and is unique. It is easy to verify that

$$A+ = V\Lambda U^T \tag{15.4}$$

where

$$A = \begin{pmatrix} \frac{1}{\sigma_1} & & & 0 \\ & \frac{1}{\sigma_2} & & \\ & & \ddots & \\ & & & \frac{1}{\sigma_r} \\ \hline 0 & & & 0 \end{pmatrix}_{n \times J} .$$

In recent years there have been a number of algorithms proposed for computing the pseudo-inverse of a matrix. These algorithms usually depend upon a knowledge of the rank of the matrix or upon some suitable chosen parameter. For example in the latter case, if one uses (15.4) to compute the pseudo-inverse, then after one has computed the singular value decomposition numerically it is necessary to determine which of the singular values are zero by testing against some tolerance.

Alternatively, suppose we know that the given matrix A can be represented as

$$A = B + \delta B$$

where δB is a matrix of perturbations and

$$\|\delta B\| \leq \eta.$$

Now, we wish to construct a matrix $\hat{B}$ such that

$$\|A - \hat{B}\| \leq \eta$$

and

$$\text{rank } (\hat{B}) = \text{minimum.}$$

This can be accomplished with the aid of the solution to problem (C). Let

$$B_k = U\Omega_k V^T$$

where Ω_k is defined as in (15.2).

Then using (15.3),

$$\hat{B} = B_p$$

if

$$(\sigma_{p+1}^2 + \sigma_{p+2}^2 + \cdots + \sigma_n^2)^{\frac{1}{2}} \leq \eta$$

and

$$(\sigma_p^2 + \sigma_{p+1}^2 + \ldots + \sigma_n^2)^{\frac{1}{2}} > \eta.$$

Since rank $(\hat{B})$ = p by construction,

$$\hat{B}^+ = V\Omega_p^+ U^T .$$

Thus, we take $\hat{B}^+$ as our approximation to A^+ .

E) Let A be a given matrix, and $\underset{\sim}{b}$ be a known vector. Determine $\hat{\underset{\sim}{x}}$ so that amongst all $\underset{\sim}{x}$ for which $\|\underset{\sim}{b} - A\underset{\sim}{x}\| = \min$, $\|\hat{\underset{\sim}{x}}\| = \min$. It is easy to verify that

$$\hat{\underset{\sim}{x}} = A^+\underset{\sim}{b} .$$

16. Calculation of the SVD.

In [14] it was shown by Golub and Kahan that it is possible to construct a sequence of orthogonal matrices

$$\left\{P^{(k)}\right\}_{k=1}^{n}, \quad \left\{Q^{(k)}\right\}_{k=1}^{n-1}$$

via Householder transformation so that

$$P^{(n)}P^{(n-1)} \cdots P^{(1)}AQ^{(1)}Q^{(2)} \cdots Q^{(n-1)} \equiv P^{T}AQ = J$$

and J is an m x n bi-diagonal matrix of the form

$$J = \begin{bmatrix} \alpha_1 & \beta_1 & 0 & \cdots & 0 \\ & \alpha_2 & \beta_2 & \cdots & 0 \\ & & \ddots & & \\ & & & \beta_{n-1} & \\ & & & \alpha_n & \\ & & 0 & & \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \\ \}(m-n)\times n \end{matrix} .$$

The singular values of J are the same as those of A. Thus if the singular value decomposition of

$$J = X\Sigma Y^{T}$$

then

$$A = PX\Sigma Y^{T}Q^{T}$$

so that

$$U = PX, \quad V = QT \quad .$$

In [16], an algorithm is given for computing the SVD of J; the algorithm is based on the highly effective algorithm of Francis [11] for computing the eigenvalues.

It is not necessary to compute the complete SVD when a vector $\underset{\sim}{b}$ is given. Since $\underset{\sim}{\hat{x}} = V\Sigma^{+}U^{T}\underset{\sim}{b}$, it is only necessary to compute V,Σ and $U^{T}\underset{\sim}{b}$; note, this has a strong flavor of principal component analysis. An ALGOL procedure for the SVD will soon be published by Golub and Reinsch and a complex FORTRAN procedure for the SVD by Businger and Golub.

17. Canonical correlations

It is well-known (cf. [1]) that in order to solve for canonical correlations,, it is necessary to solve the matrix equation

$$\begin{pmatrix} 0 & \Sigma_{12} \\ \Sigma_{21} & 0 \end{pmatrix} \begin{pmatrix} \underset{\sim}{\alpha} \\ \underset{\sim}{\beta} \end{pmatrix} = \lambda \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix} \begin{pmatrix} \underset{\sim}{\alpha} \\ \underset{\sim}{\beta} \end{pmatrix} \tag{17.1}$$

where Σ_{11} is a p x p positive definite matrix and Σ_{22} is a q x q positive definite matrix. The eigenvalues of (17.1) correspond to the canonical correlations. Since Σ_{ii} is positive definite we have

$$\Sigma_{ii} = \Gamma_i^T \Gamma_i \ .$$

A short manipulation shows we may rewrite (17.1) as

$$\begin{pmatrix} 0 & \Omega \\ \Omega^T & 0 \end{pmatrix} \begin{pmatrix} \underset{\sim}{\xi} \\ \underset{\sim}{\eta} \end{pmatrix} = \lambda \begin{pmatrix} \underset{\sim}{\xi} \\ \underset{\sim}{\eta} \end{pmatrix}$$

where $\Omega \doteq \Gamma_1^{-T} \Sigma_{12} \Gamma_2^{-1}$ (17.2)

$$\underset{\sim}{\xi} = \Gamma_1 \underset{\sim}{\alpha} \ , \ \underset{\sim}{\eta} = \Gamma_2 \underset{\sim}{\beta} \ .$$

Thus by (14.2) and (14.3), we see that the canonical correlations, λ_j, are the singular values of Ω.

Suppose we have two sets of data $X_{n \times p}$ and $Y_{n \times q}$. We assume that the mean of each variable is zero. Then

$$\hat{\Sigma}_{11} \doteq cX^TX, \ \hat{\Sigma}_{22} \doteq cY^TY, \ \hat{\Sigma}_{12} \doteq cX^TY (c > 0) \ .$$

Using the Householder algorithm described in section 8 or the Gram-Schmidt algorithm described in section 10, we may write

$$X = QR,\ Q^TQ = I_p,\ R: \triangle \quad ,$$

$$Y = PS,\ P^TP = I_q,\ S: \triangle \quad .$$

Hence by (17.2)

$$\hat{\Omega} = R^{-T}R^TQ^TPSS^{-1}$$

$$= Q^TP \quad . \tag{17.3}$$

Therefore the canonical correlations are the singular values of Q^TP. Note

$$\sigma_i(\hat{\Omega}) \leq \| \hat{\Omega} \|_2 \leq \| Q^T \|_2 \cdot \|P\|_2 \leq 1 \, .$$

A short calculation shows that $\hat{\underset{\sim}{\alpha}}_i = R^{-1}\underset{\sim}{u}_i$ and $\hat{\underset{\sim}{\beta}}_i = S^{-1}\underset{\sim}{v}_i$ where $\underset{\sim}{u}_i$ and $\underset{\sim}{v}_i$ are the i^{th} columns of U and V, respectively. This method of characterizing the canonical correlations has been observed previously (cf. [25]). An algorithm using these techniques will soon be published by Björck and Golub.

Acknowledgements

I am very pleased to acknowledge the many helpful comments made by Dr. Ake Björck and Dr. Peter Businger.

References

[1] T.W. Anderson, An Introduction to Multivariate Statistical Analysis, John Wiley, New York, 1958.

[2] A. Björck, "Solving linear least squares problems by Gram-Schmidt orthogonalization," BIT, 7(1967), 1-21.

[3] A. Björck, 'Iterative refinement of linear least squares solution I', BIT, 7(1967),257-278. 'Iterative refinement of linear least squares solutions II", BIT, 8(1968),8-30.

[4] A. Björck and G.H. Golub, 'Iterative refinement of linear least square solutions by Householder transformation", BIT, 7(1967), 322-337.

[5] P. Businger and G.H. Golub, "Linear least squares solutions by Householder transformations," Num. Math., 7(1965),269-276.

[6] C. Eckart and G. Young, 'The-approximation of one matrix by another of lower rank", Psychometrika, 1 (1936),211-218.

[7] D.K. Fadeev, V.N. Kublanovskaya, and V.N. Fadeeva, "Sur les systèmes linéarires algébriques de matrices rectanqulaires et mal-conditionnées," "Programmation en Mathématiques Numériques, Editions du Centre National de la Recherche Scientifique, Paris VII, 1968.

[8] K. Fan and A. Hoffman, 'Some metric inequalities in the space of matrices", Proc. Amer. Math. Soc.,6 (1955)111-116.

[9] G.E. Forsythe and C. Moler, Computer Solution of Linear Algebraic Systems, Prentice-Hall, Englewood Cliffs, New Jersey, 1967.

[10] G.E. Forsythe and E.G. Straus, "On best conditional matrices," Proc. Amer. Math. Soc.,6 (1955),340-345.

[11] J. Francis, "The QR transformation. A unitary analogue to the LR transformation', Comput. J.,4(1961,1962)265-271.

[12] G.H. Golub, 'Comparison of the variance of minimum variance and weighted least squares regression coefficients,' Ann. Math. Statist., 34 (1963),984-991.

[13] G.H. Golub, 'Numerical methods for solving linear least squares problems",_ Num. Math., 7(1965),206-216.

[14] G.H. Golub and W. Kahan, 'Calculating the singular values and pseudo-inverse of a matrix," J. SIAM, Numer. Anal. Ser. B, 2 (1965), 205-224.

[15] G.H. Golub and J. Wilkinson, 'Iterative refinement of least squares solution," Num. Math., 9(1966),139-148.

[16] G.H. Golub, "Least squares, singular values and matrix approximations,' Aplikace Matematiky, 13(1968),44-51.

[17] B. Green, "The orthogonal approximation of an oblique structure in factor analysis", Psychometrika, 17 (1952) 429-440.

[18] R. Hanson and C. Lawson, "Extensions and applications of the Householder algorithm for solving linear least squares problems," Jet Propulsion Laboratory, 1968.

[19] H. Hotelling, 'Some new methods in matrix calculation, Ann. Math. Statist., 14 (1943), 1-34.

[20] T. Jordan, 'Experiments on error growth associated with some linear least-squares procedures, "Math. Comp., 22 (1968), 579-588.

[21] W. Kahan, "Numerical linear algebra," Canad. Math. Bull., 9 (1966), 757-801.

[22] C. Lanczos, Linear Differential Operators, Van Nostrand, London, 1951, Chap. 3.

[23] J. Longley, 'An appraisal of least squares problems for the electronic computer from the point of view of the user,' JASA, 62 (1967), 819-841.

[24] C.B. Moler, "Iterative refinement in floating point,' J. Assoc. Comput. Mach., 14 (1967), 316-321.

[25] I. Olkin, On Distribution Problems in Multivariate Analysis, Ph.D. thesis, Univ. of North Carolina (1951).

[26] R. Penrose, 'A generalized inverse for matrices', Proc. Cambridge Philos. Soc., 51 (1955) 406-413.

[27] J. Rice, "Experiments on Gram-Schmidt Orthogonalization," Math. Comp., 20 (1966), 325-328.

[28] P. Schönemann, "A generalized solution of the orthogonal procrustes problem', Psychometrika, 31 (1966), 1-10.

[29] A. van der Sluis, 'Condition numbers and equilibration of matrices', (unpublished report from Utrecht University, ERCU 37-4).

[30] J.H. Wilkinson, RoundingErrors in Al ebrage Processes, Prentice-Hall, Englewood Cliffs, New Jersey, 19 3. 96

[31] J.H. Wilkinson,"A priori error analysis of algebraic processes', Proceedings of the International Congress of Mathematicians, Moscow, 1966, 629-640.

[32] J.H. Wilkinson, 'The solution of ill-conditional linear equations', Mathematical Methods for Digital Computers, Vol. II, A. Ralston, Ph.D. and H. Wilf, Ph.D., ed., John Wiley, New York, 1967, 65-93.